Journal of Network and Information Security
8 (1&2), 2020, 31-35
http://www.publishingindia.com/jnis

Anomalous Insiders Detection System using K-NN in Collaborative Information Systems

Thiraviaselvi G.[1*] and Dhinese G.[2]
[1]Assistant Professor, Department of Information Technology, Francis Xavier Engineering College, Tirunelveli, Tamil Nadu, India. Email: thiraviaselvi@gmail.com
[2]Web Developer, Wicmad Technologies, Tirunelveli, Tamil Nadu, India. Email: dhinese@gmail.com
*Corresponding Author

Abstract: Collaborative Information Systems (CIS) allow users to belong to different groups to communicate and interfere with shared tasks or documents for collaboration. Current Intrusion Detection Systems are not effective in detecting insider threats where users work in dynamic teams. A malicious hacker who works as an employee of an organization or an outsider who acts as an employee by obtaining false credentials is called an insider threat and that malicious hacker may cause damages to the shared information. The proposed Neighborhood Anomaly Detection System (NADS), is an unsupervised learning framework to detect insider threats. NADs makes use of access logs of collaborative environments for Intrusion Detection. This framework is based on the observation that typical CIS users tend to form Neighborhood structures based on the subjects accessed. NADS consists of two components: 1) relational pattern extraction, where Neighborhood structures are derived and 2) anomaly prediction, which uses a statistical model based on relational pattern extraction. Based on the observations, the deviation of users from the communities they belong to is detected. It is capable to detect anomalous insiders in systems that use dynamic teams.

Keywords: Anomaly detection, Data mining, Insider threat, Network analysis.

I. Introduction

The fluctuating nature of coordination among human beings is the main cause of the evolution of Collaborative Information System (CIS). For example, in an organization, a person can make decisions on his or her own. But, this situation may affect other employees of an organization. In some cases, several expertise may be pooled together to form a team. All these complexities of collaboration between team members lead to the evolution of CIS model.

Bringing marketing, manufacturing and design information together is the overall goal of CIS. So, all the employees can work towards a single goal. Since the goal of all employees is to earn a profit, it is assumed that if the appropriate information is given, the desired action will be followed.

Collaborative Information Systems (CIS) allow a group of users to interfere with shared tasks. At the same time, CIS is responsible for managing sensitive information. Unauthorized accessing of information from such environments is dangerous to both the managing agencies and individuals those who are the proprietors of that information. Greatest security threats to such CISs are insiders.

Insider threats are malicious hackers who get access to the systems with false credentials. Numerous approaches have been developed to detect insider threats in CISs. In this paper, a framework is introduced to detect insider threat from the access logs of CIS. This framework is called Neighborhood Anomaly Detection System. Here, a user must have a similarity with other users. In addition, meta-information is also included.

Many approaches have been developed to address the anomalous insiders in collaborative environments. First, access control models assume a user's role (or their relationship to a group). However, CIS often interrupt this principle because teams may be constructed dynamically, based on the everchanging needs of the environment and the availability of the users.

The greatest security threat to information systems is insiders. This framework focuses on detecting insider threats in centralized CIS which is managed by an individual organization. A suspicious insider corresponds to an authenticated user whose actions run security to the organization's policies.

II. Methodology

Many observations have been made before to detect anomalous insiders using various techniques. In Principal Component Classifier [1] method distance finding, Principal Component Analysis (PCA) and Outlier Detection are done. Dimensionality Reduction is the main advantage. PCA makes statistics less efficient. However, it provides 99% average detection rate when compared to other methods.

Parallelism [2] may detect insiders in an accurate way. Because a greater number of Anomaly Detection algorithms are implemented parallel in a system. It produces 10% lower false-positive rate. However, it also has some disadvantage (i.e.) inherent delay in obtaining anomaly detection results and large trace files are needed to be transmitted and stored. Conditional Anomaly Detection [3] detects anomaly from an outlier in a conditional way. Here, scalability is a problem, because of multi-gigabyte database usage.

Graph-based Anomaly Detection [4] uses graph theory to detect insiders, knowledge about graph theory is essential. Detection based on histogram [4] construction, used in enterprise networks may select some features and build histogram based on the features. This approach merges different normal behaviours. So, the ability to discriminate anomalous insiders from normal conditions declines. Fig. 1 represents the architectural overview of NADs and MetaNADs.

Mimicry attacks [5] can be detected from SNAD (Specialized Network Anomaly Detection) framework. But it may have the following disadvantages: 1) Not appropriate for large Networks. 2) Neglects semantics of the relation. Anomaly Detection [6] based on the behavior of users, profile mining uses data mining concept. But it has a false-positive rate.

III. Overview of Framework

This framework comprises of two primary components: 1) NADS and 2) MetaNADS. One of the challenges in working with collaborative environment access logs is they do not explicitly article the community structure of the organization. By recognizing this deficiency, NADS makes the relationship between users and subjects. MetaNADS makes the relationship between subjects and the categories to which subjects belong to.

A. NADS

NADS is one of the primary components in this framework. It has two parts of the work. 1) Access Network construction and 2) Pattern Extraction. Access Network denotes the relationship between users and subjects. The pattern is extracted from the constructed network access logs. This pattern consists of users and subjects.

i. Construction of Access Network

Consider a matrix A of size |S|×|U| where S denotes subjects and U denotes users. This matrix represents the affinity that a user has toward a particular subject when assessing the similarity of a group.

ii. Extracting Patterns

The similarity between the users is calculated from the matrix A by matching the subjects accessed by the first user to the subjects accessed by other users. Then do the same for the second user and so on. Mathematically A (i, j-1) is equalized to A (i-1, j). If it is equal, then increment the similarity value of i.

B. MetaNADS

This section begins with an overview of the MetaNADS

i. Construction of Assignment Network

Consider a matrix B of size |G||×|U| where G denotes the Category and U denotes the users. This matrix denotes the attraction that a user has toward a category when evaluating the similarity of a group.

ii. Extracting Patterns

The similarity among the users is calculated from the matrix B by matching the categories of subjects accessed by the first user to the categories of subjects accessed by another user. Then do the same for the second user and so on. Mathematically, B (i, j-1) is equalized to B (i-1, j). If it is equal, increment the similarity value of i.

C. Detection of Anomalous Insiders

This section explains Neighborhood formation and anomaly detection.

i. Formation of Communities

If even a single subject accessed by a user is similar to subjects accessed by another user then form a Neighborhood, likewise, do the same for other users.

a) Discovery of Nearest Neighbors

K-nearest neighbors (KNNs) of a user, is detected by calculating the distance among the users. Distance is the similarity among the users while accessing the subjects. The value of K is heuristics.

b) Deviation Measurement from Nearest Neighbors

The radius of a user ui is defined as the distance to its kth nearest neighbor excluding itself. Specifically, the radius of ui is ri = sort(DIS(i; :))(i; k +1), where sort is a function that ranks distances in increasing order from smallest to largest. Users are thus characterized as a vector of radius r = [r1; r2; . . . ; r|U|] and set of neighbors knn = [knn1; knn2; . . . ; knn|U|]. The density of the user's network will be higher when the radius becomes small.

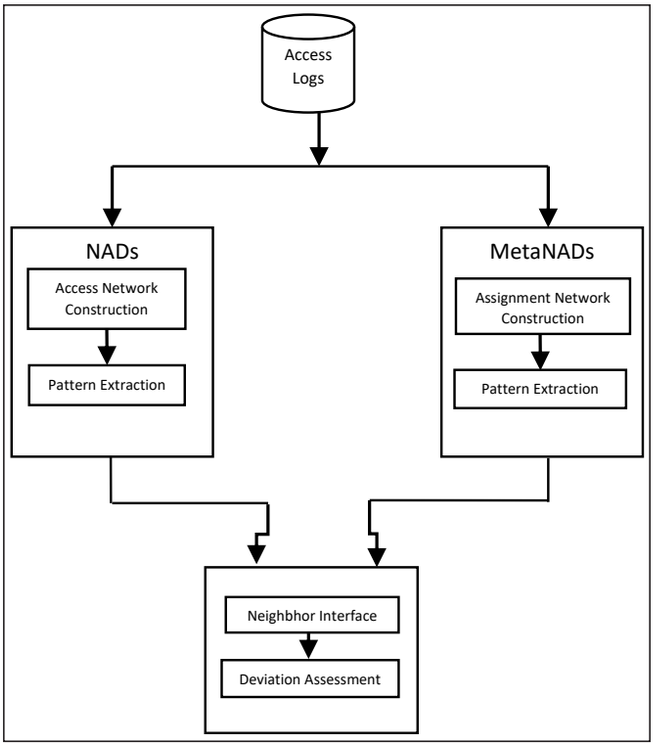

Fig. 1: Architectural Overview

Insider Threats cannot be detected through radius alone. Consider, in Fig. 2, user uy and the users in cluster F can be correctly classified as anomalous based on their radius. In contrast, we would fail to detect ux as an anomaly because it has a smaller radius in comparison to nodes in the F area. This bias is due to a reliance on raw values and thus we normalize the system. Rather than use raw radius, we calculate the deviation of a node's radius from those of its k- nearest neighbors to assess the degree to which it is anomalous.

$$Dev(u_i) = \sqrt{\frac{\sum_{u_j \in knn_i} (r_j - \bar{r})^2}{k-1}} \quad (1)$$

such that k denotes the no. of Nearest Neighbors, r denotes the Radius of the user, Knn denotes the set of neighbors, ui, uj denotes the users.

$$\bar{r} = \sum_{u_j \in knn_i} r_j /(k) \quad (2)$$

Turning back to Fig. 2, the radius deviations of the nodes in area E are much smaller than those in F, such that the deviation of node ux is much larger than uy. Normal users are likely to exhibit significantly smaller radius deviation scores than abnormal users. u1, u3, and u6 receive larger deviation scores in NADS than they do in MetaNADS.

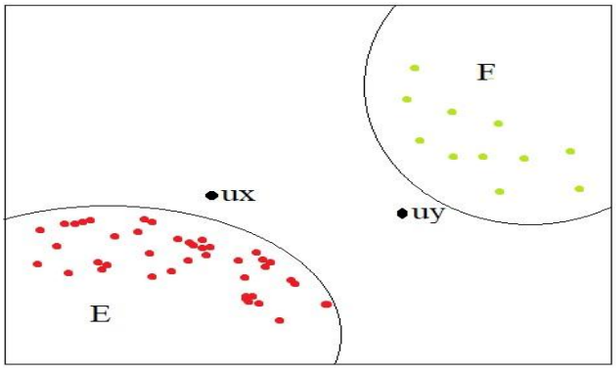

Fig. 2: An Example of the Influence of Radius Size in Nearest Neighbor Sets on Insider Threat Detection

IV. Experimental Results

In this paper, the Electronic Health Record System is implemented in Java using jdk1.6 as a tool. EHR system consists of 9 Categories and 98 Subjects are included. For discussions, 2689 communications are considered. Among the available online users, Neighborhood is formed using k-nearest neighbor discovery. The value of K is taken as 2. From the communities, an anomaly is predicted using deviation assessment.

A. EHR Access Log Dataset

EHR access logs consist of information about the subjects, category, users and the time at which the user accessed that particular subject. The user interfaces are accessible through LAN. In all, EHR stores over 2689 observations on over 98 patient records.

The access communications are divided into two parts: 1) user-subject access communications of the form <user, subject> and 2) user-category communications of the form <user, category>. There are 2689 user-subject communications and user-category communications in the analyzed data set Anomaly detection models are based on the availability of online users and the performance is reported.

B. Deviation Score Distributions

i. NADS

Fig. 3 provides an example of the deviation score distributions of NADS on an arbitrary day of accessed subjects in the EHR dataset. Fig. 4 provides an example of the number of users having the same deviation score distributions of NADS in the EHR dataset.

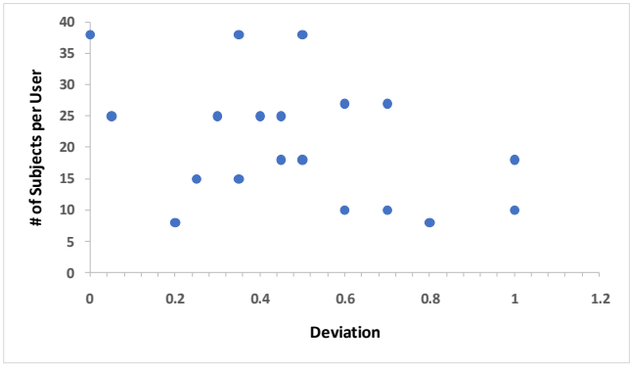

Fig. 3: Distribution of User Deviations on an Arbitrary Day

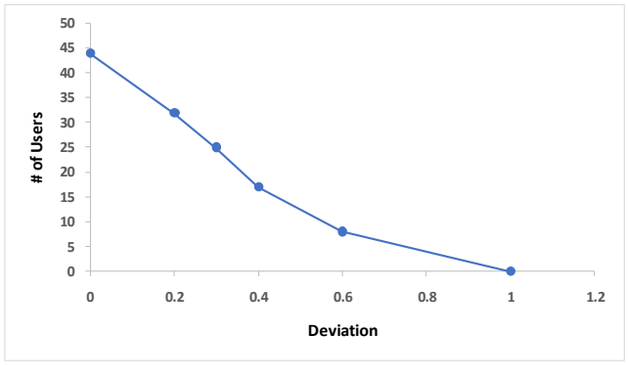

Fig. 4: Distribution of User Deviations based on the Number of Users

ii. MetaNADS

Fig. 5 provides an example of the deviation score distributions of MetaNADS on an arbitrary day. Fig. 6 provides an example of the number of users having the same deviation score distributions of MetaNADS in the EHR dataset.

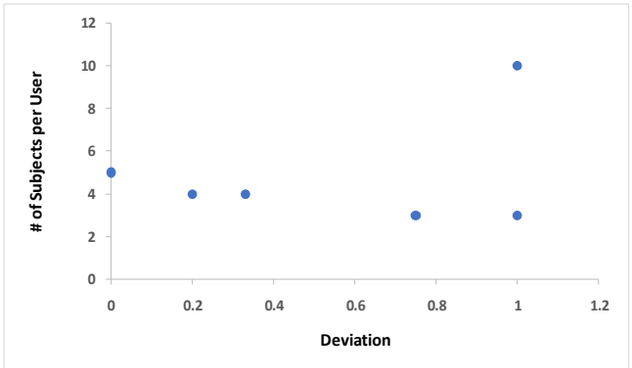

Fig. 5: Distribution of User Deviations on an Arbitrary Day

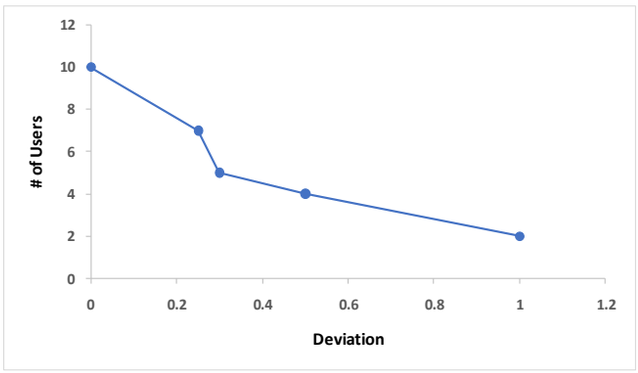

Fig. 6: Distribution of User Deviations based on the Number of Users

iii. Comparison between NADS and MetaNADS

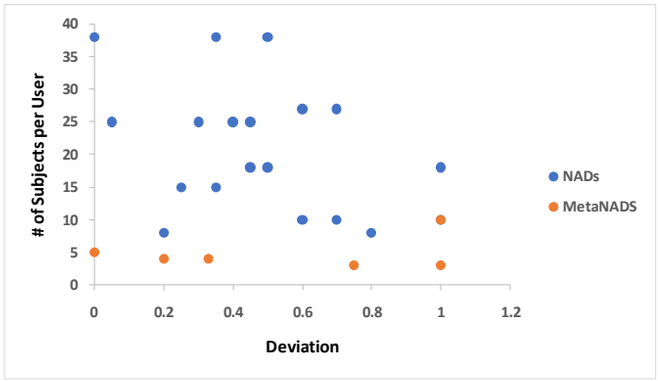

Fig. 7: Comparison between NADS and MetaNADS - Distribution of User Deviations on an Arbitrary Day

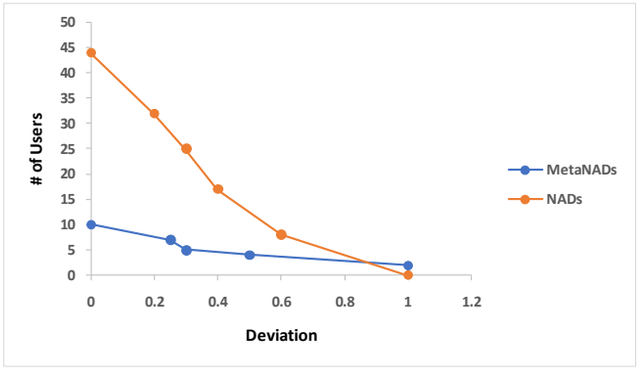

Fig. 8: Comparison between NADS and MetaNADS - Distribution of User Deviations based on the Number of Users

Fig. 7 and Fig. 8 exhibits some deviation scores, which is a result of combining the access network and assignment

network relations. This leads to larger user communities. Here, majority user deviations are relatively small. Nearly 60% of users receive a deviation score of less than 0.8.

V. Conclusion

To detect anomalous insiders in a CIS, NADS, a Neighborhood anomaly detection system that utilizes a relational framework is proposed. NADS calculates the deviation of users based on their nearest neighbor networks to predict which users are anomalous. It is found that "normal" users tend to form communities, unlike illicit insiders. NADS dominates when the number of insider threats increases but MetaNADS is the best model when the number of intruders is relatively small.

References

[1] M.-L. Shyu, S.-C. Chen, K. Sarinnapakorn, and L. Chang, "A novel anomaly detection scheme based on principal component classifier," *Third IEEE International Conference on Data Mining (ICDM'03),* 2003.

[2] S. Shanbhag, and T. Wolf, "Accurate anomaly detection through parallelism," *IEEE Network*, vol. 23, no. 1, pp. 22-28, 2009.

[3] X. Song, M. Wu, C. Jermaine, and S. Ranka, "Conditional anomaly detection," *IEEE Transactions on Knowledge and Data Engineering*, vol. 19, no. 5, pp. 631-645, 2007.

[4] A. Kind, M. P. Stoecklin, and X. A. Dimitropoulos, "Histogram-based traffic anomaly detection," *IEEE Transactions on Network and Service Management*, vol. 6, no. 2, pp. 110-121, 2009.

[5] Y. Chen, S. Nyemba, W. Zhang, and B. Malin, "Leveraging social networks to detect anomalous insider actions in collaborative environments," *Proceedings of IEEE Ninth Intelligence and Security Informatics*, pp. 119-124, 2011.

[6] W. Eberle, and L. Holder, "Applying graph-based anomaly detection approaches to the discovery of insider threats," *Proceedings of IEEE International Conference on Intelligence and Security Informatics*, pp. 206-208, 2009.